Item Response Theory analysis of the Cognitive Reflection Test: Testing the psychometric properties of the original scale and a newly developed 8-item version

Item Response Theory analysis of the Cognitive Reflection Test: Testing the psychometric properties of the original scale and a newly developed 8-item version

Caterina Primi (primi@unifi.it)
Department of NEUROFARBA – Section of Psychology, University of Florence
Via di S.Salvi 12- Padiglione 26 – 50135 Florence (Italy)

Kinga Morsanyi (k.morsanyi@qub.ac.uk)
School of Psychology, Queen's University Belfast,
Belfast, BT7 1NN, Northern Ireland, UK

Maria Anna Donati (mariaanna.donati@unifi.it) & Francesca Chiesi (francesca.chiesi@unifi.it)
Department of NEUROFARBA – Section of Psychology, University of Florence
Via di S.Salvi 12- Padiglione 26 – 50135 Florence (Italy)

Abstract

The Cognitive Reflection Test (CRT) is a short measure of a person's ability to resist intuitive response tendencies, and to produce a normative response which is based on effortful reasoning. The CRT correlates strongly with important real-life outcomes, such as time preferences, risk-taking, and rational thinking. Although the CRT is a very popular measure, there is virtually no available data about its psychometric properties. The present study aimed at investigating the psychometric properties of the CRT, and to verify the suitability of a longer version of the test, which was obtained by adding five new items to the three original ones. We applied Item Response Theory analyses. The two-parameter logistic model was used in order to estimate item parameters (difficulty and discrimination), and the Test Information Function was computed to assess the measurement precision of both the original and the longer versions of the test. The results confirmed the suitability of the original items for measuring the cognitive reflection ability trait. Furthermore, the results demonstrated that the longer version of the scale measures with high precision a wider range of the cognitive reflection latent trait.

Keywords: cognitive reflection; individual differences; item response theory; test information function.

Introduction

The Cognitive Reflection Test (CRT; Frederick, 2005) is a short test measuring a person's tendency to override an intuitively compelling response, and to engage in further reflection which can lead to a correct solution. As an example, consider the following item: *A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? _____ cents.* Although the correct response is 5 cents, many participants give the response "10 cents", which seems to pop into mind effortlessly. Indeed, a remarkable property of the CRT is that for each item, almost all participants produce either the normatively correct response, or the same incorrect (i.e., heuristic) response. That is, reasoning errors are very systematic.

It has been proposed that because the typical heuristic response comes very quickly and easily (i.e., fluently) to mind, people will be highly confident that this answer is correct, and will be reluctant to revise it (cf. Thompson & Morsanyi, 2012). Indeed, in a subsequent study, De Neys, Rossi and Houdé (2013) found that people who gave the incorrect response to the bat and ball problem were 83% confident that their response was correct. Although this was significantly lower than the 93% confidence level reported by the participants who gave the correct response, this still demonstrates the attractiveness of the heuristic response. That is, to be able to produce a correct response, participants have to be able to effectively monitor and correct their impulsive response tendencies (cf., Frederick, 2005).

Cognitive reflection was found to be negatively related to temporal discounting (i.e., the tendency to prefer smaller, immediately available rewards over larger rewards which will be available later), and positively related to choosing gambles with higher expected values (Frederick, 2005). Further studies showed that the CRT was also related to some typical heuristics and biases (e.g., Liberali, Reyna et al., 2012; Toplak, West & Stanovich, 2011, 2013), including tasks that contained no numerical information (such as syllogistic reasoning problems). Furthermore, although the CRT correlates with measures of intelligence and numeracy (e.g., Frederick, 2005), it was found to explain additional variance in reasoning and decision-making tasks when it was administered together with measures of intelligence and numeracy (Liberali et al., 2012; Toplak et al., 2011). Other studies showed an association between the CRT and metacognitive skills (Mata, Ferreira & Sherman, 2013), and people's motivation to fully understand causal mechanisms (Fernbach, Sloman, Louis & Shube, 2013), and a negative association between the CRT and superstitious and paranormal beliefs (Pennycook, Cheyne,

Seli, Koehler & Fugelsang, 2012). Overall, these results demonstrate that the CRT is a very powerful predictor of a person's ability to make unbiased judgments and rational decisions in a wide variety of contexts. However, as a consequence of the huge popularity of the CRT, the three original items have become extremely well-known. This obviously weakens the suitability of the original scale in measuring cognitive reflection, as participants might know the correct responses already.

A further issue is the difficulty of the original items. Indeed, in his original publication, Frederick (2005) reported that in some university student samples, more than 50% of the respondents scored 0 on the test. Thus, the test might not be suitable for lower ability or less educated samples. Finally, with only three items, it is necessarily difficult to discriminate with high precision between people with different levels of cognitive reflection.

Given these issues regarding the CRT, the aim of the present study was to develop some new items with similar characteristics to the original ones, in order to create a new version of the test, which is at least partially unknown to participants. In developing this longer version of the CRT, we started by investigating the psychometric properties of the original problems, since despite the widespread use of the CRT its psychometric properties are virtually unknown. In his original publication, Frederick (2005) did not report the reliability of the scale, and, with a few exceptions (Campitelli & Gerrans, 2013; Liberali et al., 2012; Weller et al., 2012), most researchers who used the scale followed the same practice. Very recently, Toplak, West & Stanovich (2013) also developed a longer version of the scale. However, this was based on a single study with a relatively small sample of participants (n=160), and the psychometric properties of the scale were not adequately tested. These authors also did not demonstrate that their participants mainly generated either the heuristic or the correct response when they responded to the new items. Finally, one of Toplak et al.'s proposed item was not open ended, but participants had to choose from three response options, which is different from the format used in the original CRT.

In the present work, we applied Item Response Theory (IRT). We chose IRT since its application have potential benefits in testing and improving the accuracy of assessment instruments. Indeed, IRT models provide item parameters that enable the evaluation of how well an item performs in measuring the underlying construct. More specifically, IRT is a model that provides a linkage between item responses and the latent characteristic assessed by a scale. IRT assumes that each examinee responding to a test item possesses some amount of the underlying ability (denoted by the Greek letter theta). It is assumed that, whatever the ability, it can be measured on an arbitrary underlying ability scale having a midpoint of zero, a unit of measurement of one, and a range from negative infinity to positive infinity (practical considerations usually limit the range of values from, say, -3 to +3).

At each ability level, there will be a certain probability that an examinee will give a correct response to the item. This probability will be denoted by $P(\theta)$. If one plotted $P(\theta)$ as a function of ability, the result would be a smooth S-shaped curve (see Figure 1). This S-shaped curve, known as the Item Characteristic Curve (ICC), describes the probability of correct response to an item as a function of the possessed ability. This probability will be small for examinees with low ability and large for examinees with high ability. The probability of a correct response is near zero at the lowest levels of ability. It continues to increase up to the highest levels of ability, where the probability of producing a correct response approaches 1.

Each item in a test will have its own item characteristic curve, depending on its specific properties. Thus, IRT attempts to model the relationship between an observed variable and the probability that the examinee will correctly respond to a particular test item. Although a number of different IRT models exist, the most commonly employed one is the two-parameter logistic model (2PL), which assumes a single underlying ability and two item parameters: a difficulty parameter (b) and a discrimination parameter (a). Measures of model fit and parameter estimates are obtained through maximum likelihood estimation.

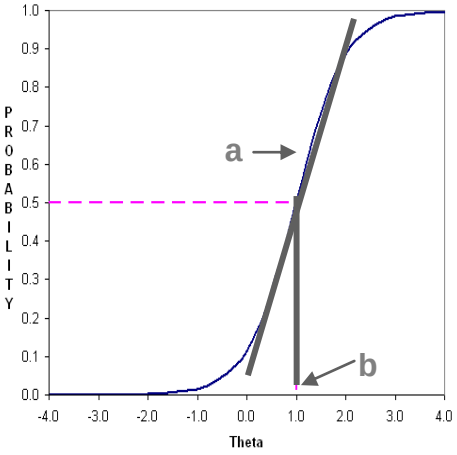


Figure 1. Exemplar Theoretical Item Characteristic Curve for the Two-Parameter Logistic Item Response Theory Model (a = discrimination, b = difficulty).

As we described above, IRT derives the probability of each response as a function of some item parameters. In the 2PL model, the first one is the difficulty (b) of the item. Under IRT, the difficulty of an item describes where the item functions along the trait, and it can be interpreted as a location index with regard to the trait being measured. For example, a less difficult item functions among the low-trait respondents and a more difficult item functions among the high-trait respondents. The second item property is discrimination (a), which describes how

well an item can differentiate between examinees with different levels of ability. The slope corresponds to item discrimination. It describes how rapidly the probabilities change in correspondence with changes in ability levels. This property is essentially reflected by the steepness of the item characteristic curve. The steeper the curve, the better the item can discriminate between levels of ability. The flatter the curve, the less the item is able to discriminate.

Additionally, IRT makes it possible to assess the measurement precision of the test through the *Test Information Function* (TIF), which, instead of providing a single value (e.g., coefficient alpha) for reliability, evaluates the precision of the test at different levels of the measured construct (Embreston & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). The information function is the expected value of the inverse of the error variances for each estimated value of the underlying construct [$I(\theta) \approx 1/SE^2(\theta)$]. This means that the more information provided by a test at a particular ability level, the smaller the errors associated with ability estimation. The Test Information curve shows graphically how well the construct is measured at different levels of the underlying construct continuum (i.e., a rather flat curve indicates that the test is discriminating within a broad range of ability; a peak means that the test is reliable in a narrow region of the latent trait distribution).

In sum, in the present work, using IRT we analyzed the properties of the original CRT items and the possibility to obtain a longer scale with novel items that are unknown to participants, and, thus, it is not possible for them to retrieve the correct responses from memory. Additionally, we expected to obtain a longer scale which measures with high precision a wider range of the cognitive reflection ability trait. Finally, we studied the validity of both the original CRT scale and the new, longer scale. In particular, we expected that the longer scale would show similar correlations with various measures of intelligence, numeracy, and decision making as the original CRT.

Methods

Participants

The participants were 988 students (Mean age = 20.2 years; *SD* = 1.8; 63% female; 76% Italian and 24% British) attending the senior year of high school (40%) and the first or second year of university (60%) at the School of Psychology and the School of Medicine in Florence (Italy) and Belfast (United Kingdom)[1]. All students participated on a voluntary basis.

[1] Preliminarily, factor analyses were conducted separately on the Italian and English sample to check for equivalence. Results attested that the Italian and English version of the scale shared the same one-factor structure and similar patterns of factor loadings. This made it possible to merge the data to perform the subsequent unidimensional IRT analyses.

Materials

Cognitive Reflection Test - Long (CRT-L): The long version is composed of the three original items (Frederick, 2005) and five new items. The development of this 8-item scale followed several iterations of testing, item elimination and modification, using different samples of participants. We started this process using a 10-item long version of the CRT, which included the three original items, three items developed by Shane Frederick (personal communication, July 2012), one additional item, based on Van Dooren, De Bock, Hessels, Janssens and Verschaffel (2005) and three items developed by us. During the item-development process, we took into consideration the fundamental attribute of the original CRT items: that the vast majority of participants either generate the correct response, or they generate a typical incorrect (i.e., heuristic) response. For this reason, two items developed by Frederick and the item developed by Van Dooren et al.'s, were modified in order to strengthen this item characteristic. For example, Frederick's item " *If you flipped a fair coin 3 times, what is the probability that it would land "heads" at least once?*" was modified to "*If you flipped a fair coin twice, what is the probability that it would land "heads" at least once?".* In sum, the final version of our long CRT scale included the three original CRT items, one item developed by Frederick, two items, which were modified versions of items developed by Frederick, a modified version of Van Dooren et al.'s (2005) problem, and an additional item developed by us (for more details, see Primi, Morsanyi, Donati & Chiesi, submitted).

Set I of the *Advanced Progressive Matrices* (APM-Set I; Raven, 1962) is a measure of fluid intelligence, and it was used as a short form of the Raven's *Standard Progressive Matrices* (SPM, Raven, 1941). Set I of the APM is composed of 12 items with increasing levels of difficulty, which cover the full range of difficulty of the items included in the SPM (Raven, 1962). Using IRT analysis procedures, the short form of the SPM has been found to have high reliability and validity (Chiesi, Ciancaleoni, Galli, & Primi, 2012).

The *Numeracy Scale* (NS; Lipkus, Samsa & Rimer, 2001) is composed of 11 items that assess basic probability and mathematical concepts including simple mathematical operations on risk magnitudes using percentages and proportions. A single composite score was computed based on the sum of correct responses.

The *Risk Seeking Behaviour Scale* (RSB) was composed of 8 items adapted from Frederick (2005). For each item participants indicated their preference between a certain gain and some probability of a larger gain. A composite score was created by summing these 8 items. A higher score indicated a preference for risk in order to obtain a larger amount of money.

Procedure

Participants individually completed the CRT scale in a self-administered format in the classroom. The average administration time was 15 minutes. A subsample (N=201) was also administered the APM-Set I, the NS, and the RSB. Total administration time for this subsample was one hour. For all these tests, answers were collected in a paper-and-pencil format. Each test was briefly introduced to the students and instructions for completion were given.

Results

As a preliminary step, item descriptives were calculated to check if participants, as expected, mostly generated either the correct or the typical heuristic response for each item (see Table 1).

Table 1. Percentages of correct and heuristic responses, standardized factor loadings, fit statistics, and parameters for each item of the CRT-L.

Item	*% C(H)*	*λ*	*S-χ²(df)*	*p*	*b (SE)*	*a (SE)*
1	39 (48)	.70	9.26 (6)	.16	0.38 (.06)	1.73 (.16)
2	45 (47)	.75	7.12 (6)	.31	0.18 (.05)	1.79 (.16)
3	57 (34)	.84	3.19 (5)	.67	-0.20 (.04)	2.90 (.32)
4	12 (69)	.67	15.33 (5)	.01	1.69 (.13)	1.67 (.20)
5	83 (15)	.74	6.92 (5)	.23	-1.25 (.09)	2.00 (.24)
6	56 (25)	.49	12.87 (6)	.04	-0.48 (.10)	0.83 (.10)
7	39 (35)	.76	5.00 (6)	.54	0.39 (.06)	1.89 (.18)
8	54 (33)	.75	6.46 (6)	.37	-0.09 (.05)	1.78 (.17)

Note. % represents the percentage of correct (C) and heuristic (H) responses. Standardized factor loadings λ are all significant at p = .001. Parameters were computed under the 2PL model (a = discrimination, b = difficulty). Due to the large sample size (N = 988) α was fixed at .01.

Then, the factorial structure of the CRT-L was tested through categorical weighted least squares confirmatory factor analyses implemented in the Mplus software (Muthén & Muthén, 2004). The CFI and the TLI both equalled to .98, and the RMSEA was .05, indicating a good fit (Schermelleh-Engel & Moosbrugger, 2003). Factor loadings were all significant (p < .001), ranging from .49 to .84 (see Table 1).

Having verified that a single continuous construct accounted for the covariation between item responses, unidimensional IRT analyses were performed. The 2PL model was tested in order to estimate the item difficulty and discrimination parameters. Parameters were estimated by employing the marginal maximum likelihood (MML) estimation method with the EM algorithm (Bock & Aitkin, 1981) implemented in the IRTPRO software (Cai, Thissen, & du Toit, 2011). In order to test the adequacy of the model, the fit of each item under the 2PL model was tested computing the *S-χ2* statistics. Given that using larger samples results in a greater likelihood of significant chi-square differences, the critical value of .01 rather than the usual critical value of .05 was employed (Stone & Zhang, 2003). Each item had a non-significant S-$\chi 2$ value, indicating that all items fit under the 2PL model. Concerning the difficulty parameters (*b)*, the original CRT items (1, 2 and 3) had medium level of difficulty from -.0.20 ± 0.04 to 0.38 ± 0.06 logit[2] across the continuum of the latent trait and the new items from -.1.25 ± 0.09 to 1.69 ± 0.13. With regard to the discrimination parameters (*a)*, following Baker's (2001) criteria, the original CRT items, as well as four out of the five new items had high discriminative power (*a* values over 1.34). Only item 6 had a medium (*a* < 1.34) discriminative power (see Table 1).

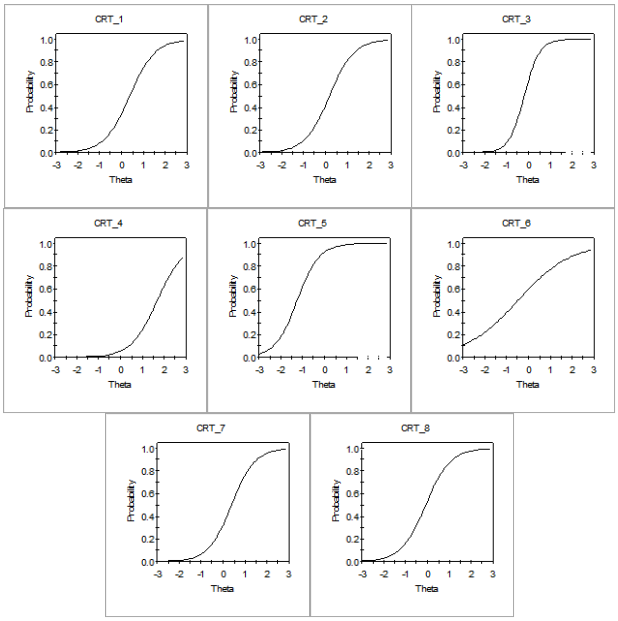
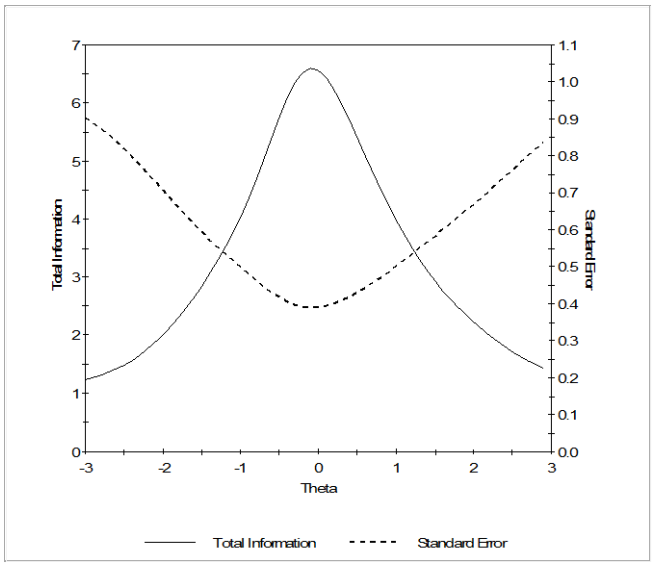


Figure 2. The ICCs of the CRT–L including the three original items (1,2 and 3) and the new items (4, 5, 6, 7, and 8) under the 2PL. Latent trait (Theta) is shown on the horizontal axis and the probability of correct responding is shown on the vertical axis.

In Figure 2 the item characteristics curves provide visual information of the item characteristics. The original CRT items were located at medium level of the trait and had a high slope, indicating high discriminative power. Concerning the new items, item 4 was located in the positive range of the trait, so it was able to measure the higher level of the trait, while item 5 was located in the negative range, so it functioned better at lower levels of the trait. All the other items had a medium level of difficulty. The slope of the new items indicated their ability to distinguish between respondents with different levels of the trait around their location.

[2] The logit is the logarithm of the *odd*, that is, the ratio between the probability of producing a correct response and the probability of responding incorrectly.



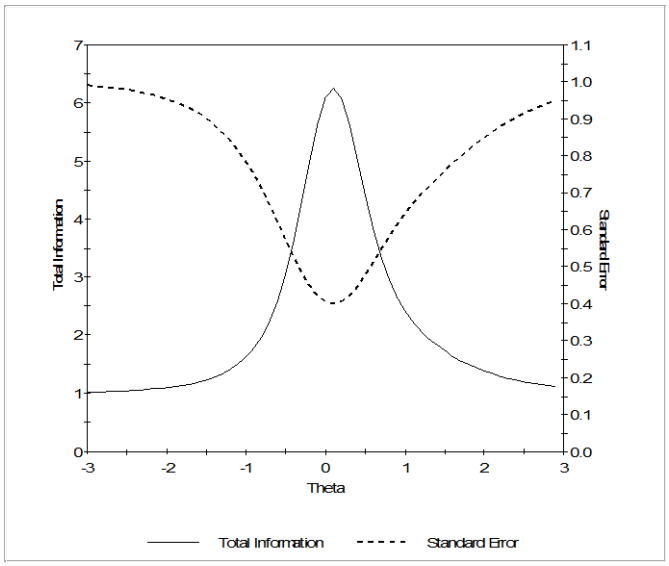


Figure 3.The Test Information Function of the CRT-L (left) and the Test Information Function including only the three original items (right) under the 2PL model. Latent trait (Theta) is shown on the horizontal axis, and the amount of information and the standard error yielded by the test at any trait level are shown on the vertical axis.

Finally, in order to identify the level of ability that is accurately assessed by the scale, the Test Information Function (TIF) was analyzed. The TIF of the CRT-L showed that the scale was informative within the range of trait from -1.20 to 1.20 standard deviations around the mean (fixed by default to 0), and the amount of information was >4 indicating that the scale was sufficiently informative (see Figure 3). Taking into consideration only the original CRT items, the TIF showed that the scale was sufficiently informative for the middle level of the trait, within the range of trait from 0.50 to 0.50 standard deviations around the mean. Comparing the two TIFs it can be seen that the CRT-L's curve has high information values associated with a larger range of the measured construct. Thus, the eight items of the CRT-L seem capable of differentiating from low-to-medium to medium-to-high levels of the latent trait, and, as such, the CRT-L allows for a better assessment of individual differences in the cognitive reflection construct than the original CRT.

Concerning validity measures, Pearson product-moment correlations attested that all the investigated relationships were significant. Regarding intelligence, both CRT scales were positively correlated with the APM (see Table 2), which is in line with previous studies (Frederick, 2005; Toplak, West & Stanovich, 2011). Concerning numeracy, we obtained a positive correlation with both CRT measures, and values appear to be similar to the values reported in previous studies employing the CRT (Cokely & Kelly, 2009; Liberali et al. 2011; Weller et al., 2013). Additionally, concerning decision making measures, Frederick (2005) observed that the original CRT was positively related to choices in risky choice tasks. That is, high CRT scores are related with more risky choices than low CRT scores. Our results are in line with this result confirming a positive correlation between risky behavior and the CRT tests.

Table 2. Correlations of CRT and CRT-L with intelligence (APM), numeracy (NS), and risk taking (RSB).

	CRT	**CRT-L**
APM	.32**	.29**
	(N=201)	(N=201)
NS	.39**	.47**
	(N=201)	(N=201)
RSB	.18*	.16*
	(N=199)	(N=199)

* $p<.05$; ** $p<.01$

Discussion

In this study we applied IRT analyses to verify the properties of a longer version of the CRT, obtained by adding five new items to the three original ones. Our analyses demonstrated that the new items had high discriminative power, similarly to the original CRT items. Moreover the five new items were more distributed along the ability scale, while the three original items were all around the mean. Thus, the TIFs showed that the new scale accurately measured a wider range of the cognitive reflection trait. In sum, these analyses confirmed the suitability of the original items in measuring cognitive reflection and also demonstrated that the new 8-item version of the scale had higher precision in measuring cognitive reflection than the original CRT. Concerning validity, the CRT-L showed similar correlations with various measures of intelligence, numeracy, and risk taking as the original CRT.

In summary, the CRT-L has the advantage of including new items that participants are unfamiliar with and offers more precision in measuring the trait across a wider range of the measured construct.

Acknowledgments

This project was supported by a British Academy/ Leverhulme Small Research Grant to K. M. and C. P. (Grant reference number: SG 120948). A longer version of this paper is submitted to the *Journal of Behavioural Decision Making*.

References

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation. Retrieved from http://info.worldbank.org/etools/docs/library/117765/Item%20Response%20Theory%20-%20F%20Baker.pdf

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–458.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO 2.1 for Windows. Chicago, IL: Scientific Software International.

Campitelli, G., & Gerrans, P. (2013). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & cognition*, 1-14. DOI: 10.3758/s13421-013-0367-9

Chiesi, F., Ciancaleoni, M., Galli, S., & Primi, C. (2012). Using the Advanced Progressive Matrices (Set I) to Assess Fluid Ability in a Short Time Frame: An Item Response Theory–Based Analysis. *Psychological Assessment*. DOI:10.1037/a0027830.

Cokely, E.T., & Kelly, C.M. (2009). Cognitive abilities and superior decision making under risk; A protocol analysis and process model evaluation. *Judgement and Decision Making*, 4, 20-33.

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*, 269-273.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fernbach, P. M., Sloman, S. A., Louis, R. S., & Shube, J. N. (2013). Explanation fiends and foes: How mechanistic detail determines understanding and preference. *Journal of Consumer Research, 39*, 1115–1131.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19,* 25–42.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M. & Pardo, S. T. (2011). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*. DOI: 10.1002/bdm.752

Lipkus,I. M., Samsa, G. & Rimer, B.K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, 21, 37-44.

Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of Personality and Social Psychology, 105*, 353–373.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus: The comprehensive modeling program for applied researchers. User's guide* (3rd ed.). LosAngeles, CA: Muthén & Muthén.

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition, 123*, 335–346.

Primi,C., Morsanyi, K., Donati, M. & Chiesi, F. (submitted). Development and testing of a new version of the Cognitive Reflection Test applying Item Response Theory (IRT). *Journal of Behavioural Decision Making.*

Schermelleh-Engel, K., & Moosbrugger, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness of fit measures. *Methods of Psychological Research Online, 8,* 23–74.

Raven, J. C. (1941). Standardization of progressive matrices. *British Journal of Medical Psychology, 19,* 137–150.

Raven, J. C. (1962). Advanced progressive matrices. London: Lewis & Co. Ltd.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40,* 331–352.

Thompson, V.A. & Morsanyi, K. (2012). Analytic thinking: Do you feel like it? *Mind & Society*, *11*, 93-105.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics and biases tasks. *Memory & Cognition, 39,* 1275-1289.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning,* DOI: 10.1080/13546783.2013.845605.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2012). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making, 26,* 198–212.

Van Dooren, W., De Bock, D., Hessels, A., Janssens, D., Verschaffel, L. (2005). Remedying secondary school students' illusion of linearity: developing and evaluating a powerful learning environment. In: Verschaffel L., e.a. (Eds.), Powerful environments for promoting deep conceptual and strategic learning (Studia paedagogica, 41) (pp. 115-132). Leuven: Universitaire Pers.